Central Analysis Facility

Mark Neubauer
Massachusetts Institute of Technology
for the CDF CAF Development Team

- **System Overview**
- **Milestones/Performance**
- **How do I get my physics done with it?**
- **Future Plans**
- **Conclusions**

The CDF CAF Group

MIT: T.Kim, M. Neubauer, F. Wurthwein

FNAL CD: R. Columbo, G. Cooper, R. Harris, R. Jetton, A. Kreymer, I. Mandrichenko, L. Weems

INFN Italy: S. Belforte, M. Casarsa, S. Giagu, O. Pinazza, F. Semaria, I. Sfligoi, A. Sidoti

Pittsburgh: J. Boudreau, Y. Gotra

Rutgers: F. Ratnikov

Carnegie Mellon: M. Paulini

Rochester: K. McFarland

Special thanks to:
- C.Paus for organizing data server population
- All ProtoCAF/Stage1 commissioning users for much needed feedback!

CAF Computing Model

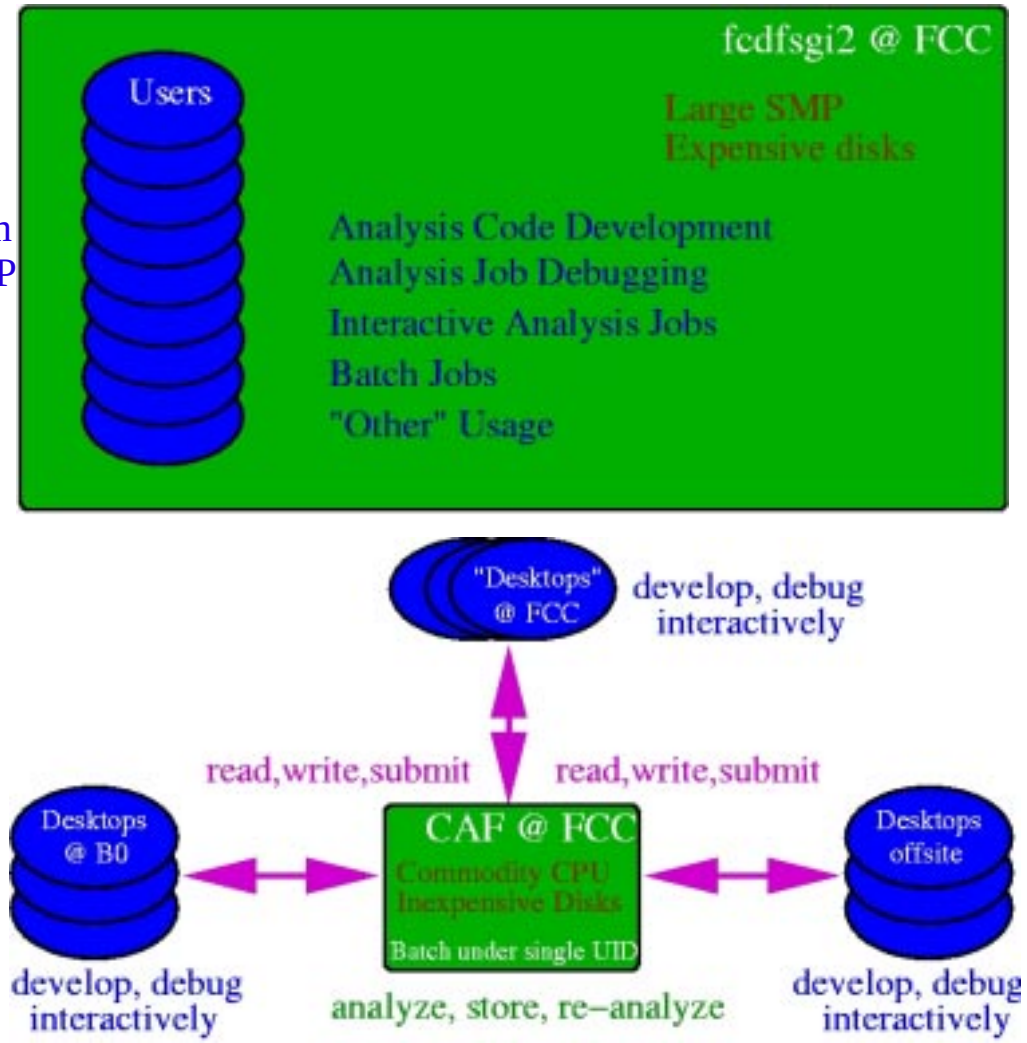
CAF design considerations:

- **Submit jobs from 'anywhere'**
- **Job output can be:**
  - sent directly to desktop
  - stored on CAF for later retrieval or input to subsequent job
- **Analysis of 5nb dataset in ~1 day for each of 200 users**
  - Requires 1THz/fb^{-1}→ cheap CPU
- **Physics groups want 150TB/fb^{-1}→need cheap disks**
  - Replace ~3 IDE drives/week

  → need hot-swap/fault-tol RAID!

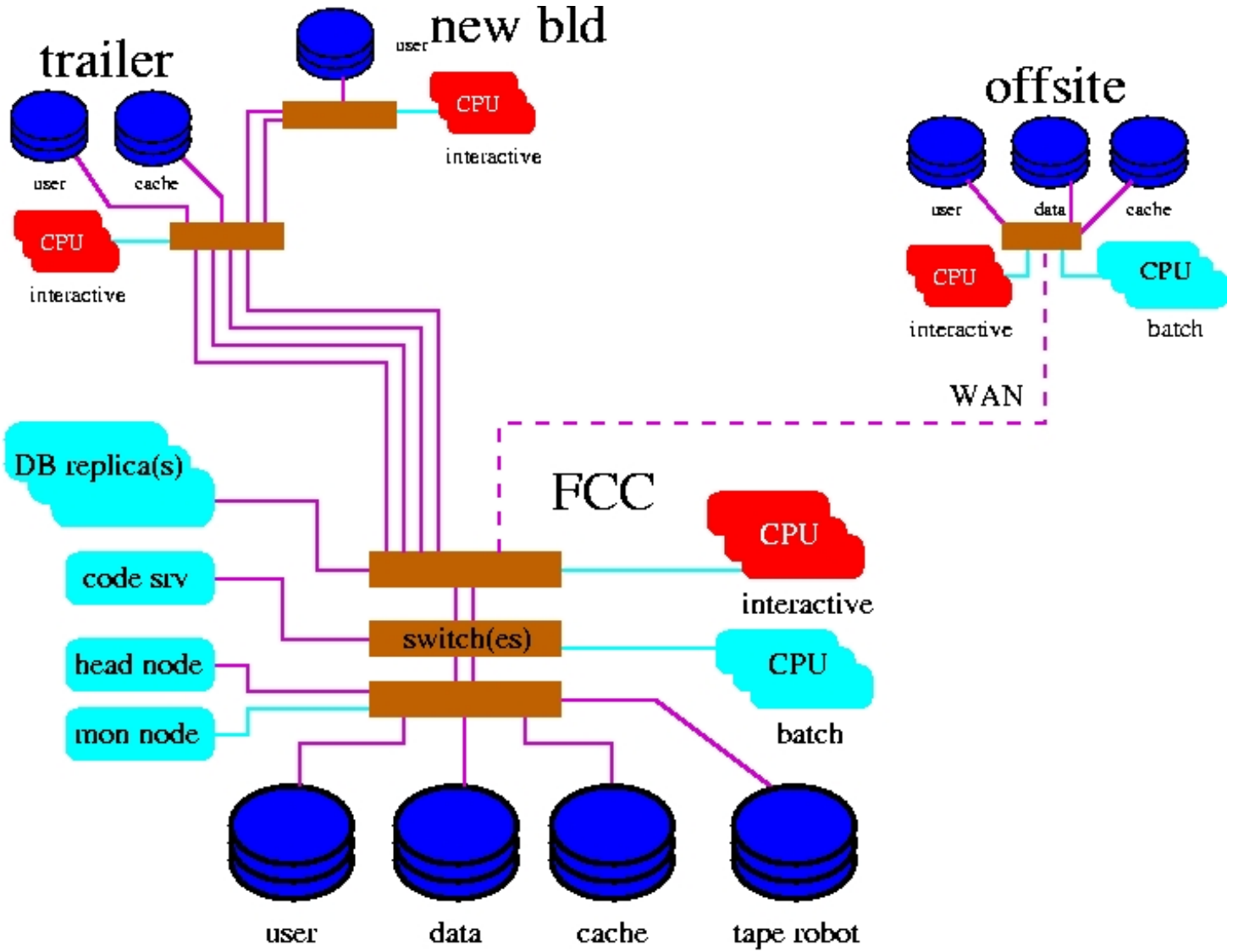
Shift from large SMP

To large PC farm



See also CDF 5743, 5787, 5802, 5914, 5961

CAF System Implementation



CAF Milestones

- CDF Central Analysis Computing Review — 8/01-11/01
- CAF prototype (protoCAF) assembled — 2/25/02
- Fully-functional prototype system (>99% job success) — 3/6/02
- ProtoCAF integrated into Stage1 system in FCC — 4/25/02
- File additional file servers (12TB total) installed — 5/3/02
- Production Stage1 CAF for collaboration — 5/30/02

ProtoCAF

Stage1

CAF Stage 1 Hardware

Workers (114 CPUs, 2U rackmount):

- 16 Dual Athelon 1.6GHz / 512MB RAM
- 41 Dual P3 1.26GHz / 2GB RAM
- FE (11 MB/s) / 80GB job scratch each

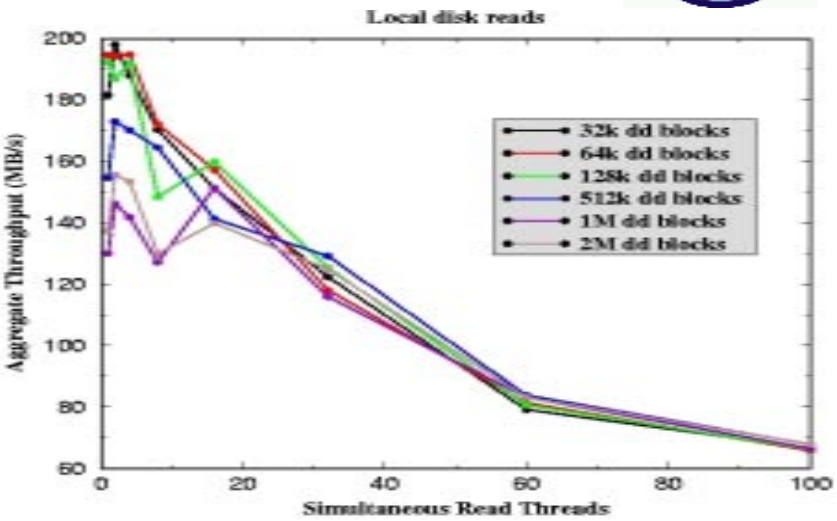
Servers (35TB total, 16 4U rackmount):

- 2.2TB useable IDE RAID50 hot-swap
- Dual P3 1.4GHz / 2GB RAM
- 1 SysKonnect 9843 Gigabt Ethernet card

Server/Client Performance:

- Up to 200MB/s local reads
- ~66MB/s NFS for 34 clients (server CPU limited)

For server evaluation details see:
http://mit.fnal.gov/~msn/cdf/caf/server_evaluation.html
or CDF 5962 (in preparation)

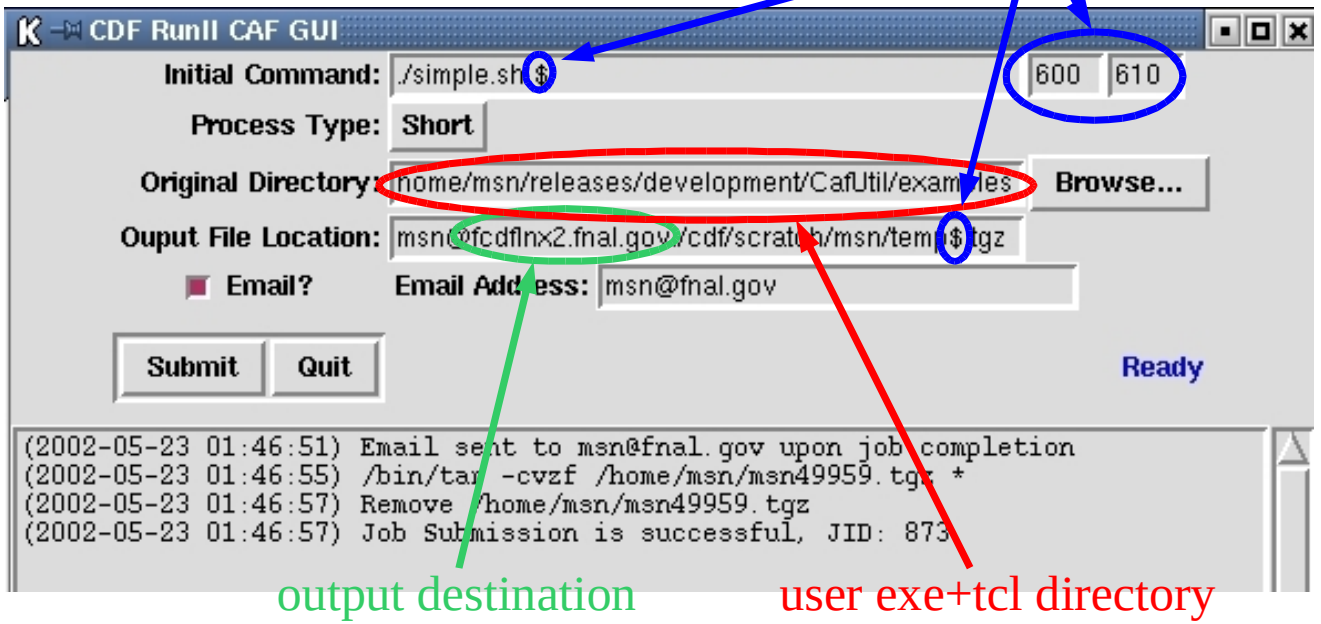


Using the CAF

- Compile, build, debug exe+tcl on 'desktop'
- Adapt CAF shell script example(s) for your job

- Enter appropriate fields & submit job via CAF GUI

CLUI also available



- Monitor job progress or just wait for email from CAF
- Retrieve output using ICAF

  ... or write output directly to 'desktop'!

Web Monitoring of User Queues

Each user a different queue

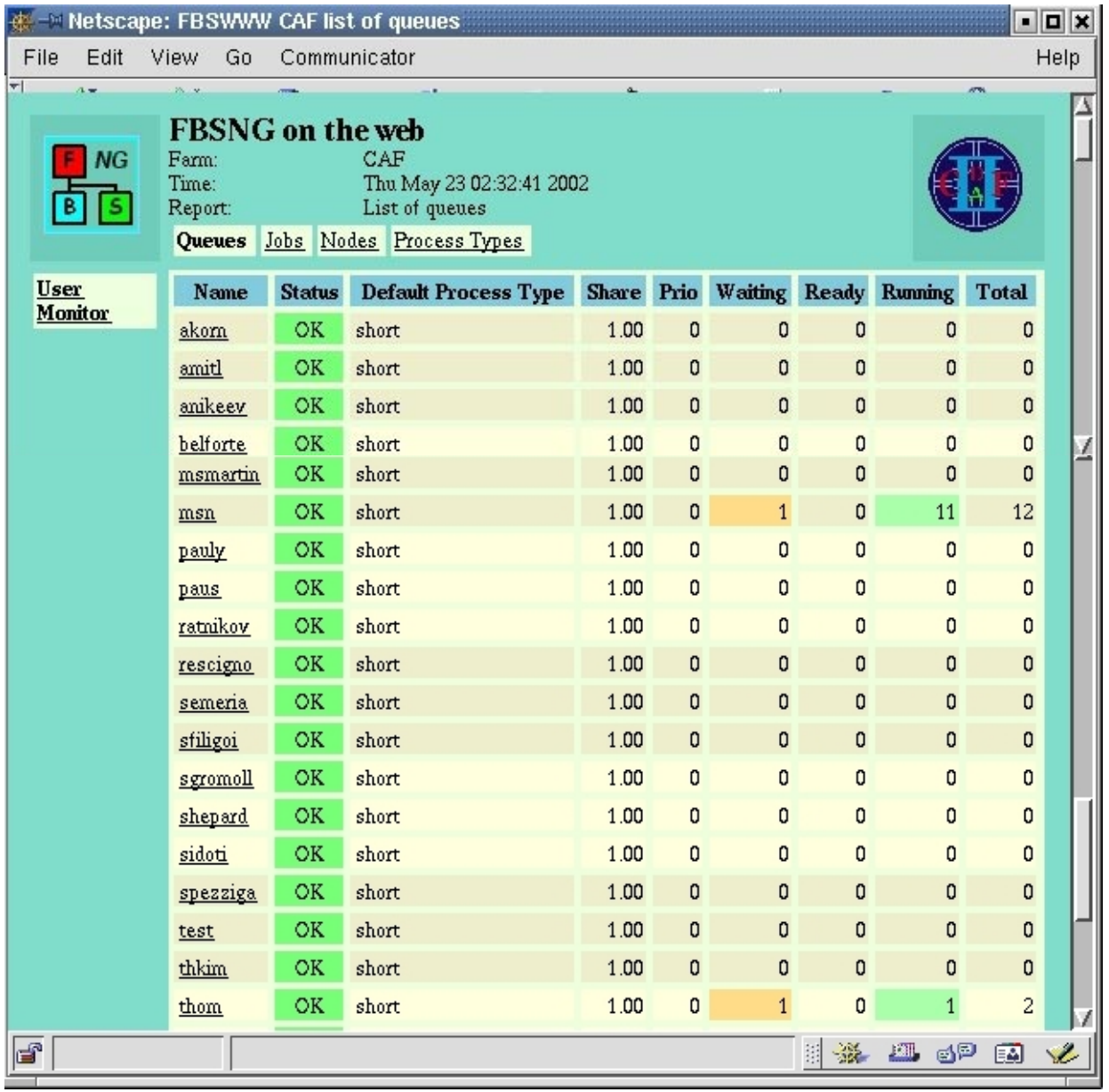
Process type for job length

- test: 5 mins
- short: 2 hrs
- medium: 6 hrs
- long: 2 days

This example:

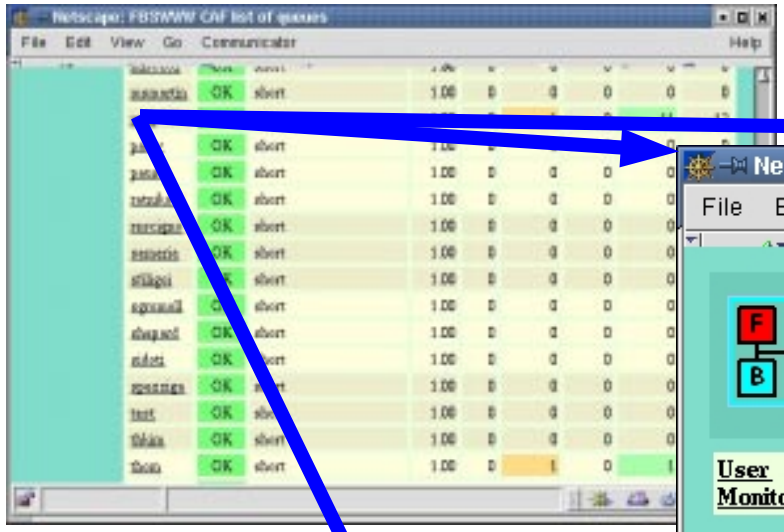
1 job→11 sections

(+ 1 additional section automatic for job cleanup)



Name	Status	Default Process Type	Share	Prio	Waiting	Ready	Running	Total
akorn	OK	short	1.00	0	0	0	0	0
amitl	OK	short	1.00	0	0	0	0	0
anikeev	OK	short	1.00	0	0	0	0	0
belforte	OK	short	1.00	0	0	0	0	0
msmartin	OK	short	1.00	0	0	0	0	0
msn	OK	short	1.00	0	1	0	11	12
pauly	OK	short	1.00	0	0	0	0	0
paus	OK	short	1.00	0	0	0	0	0
ratnikov	OK	short	1.00	0	0	0	0	0
rescigno	OK	short	1.00	0	0	0	0	0
semeria	OK	short	1.00	0	0	0	0	0
sfiligoi	OK	short	1.00	0	0	0	0	0
sgromoll	OK	short	1.00	0	0	0	0	0
shepard	OK	short	1.00	0	0	0	0	0
sidoti	OK	short	1.00	0	0	0	0	0
spezziga	OK	short	1.00	0	0	0	0	0
test	OK	short	1.00	0	0	0	0	0
thkim	OK	short	1.00	0	0	0	0	0
thom	OK	short	1.00	0	1	0	1	2

Monitoring jobs in your queue



Netscape: FBSWWW – queue msn@CAF

File Edit View Go Communicator Help

FBSNG on the web

Farm: CAF
Time: Thu May 23 01:47:23 2002
Report: Queue msn

Queues | Jobs | Nodes | Process Types

User Monitor

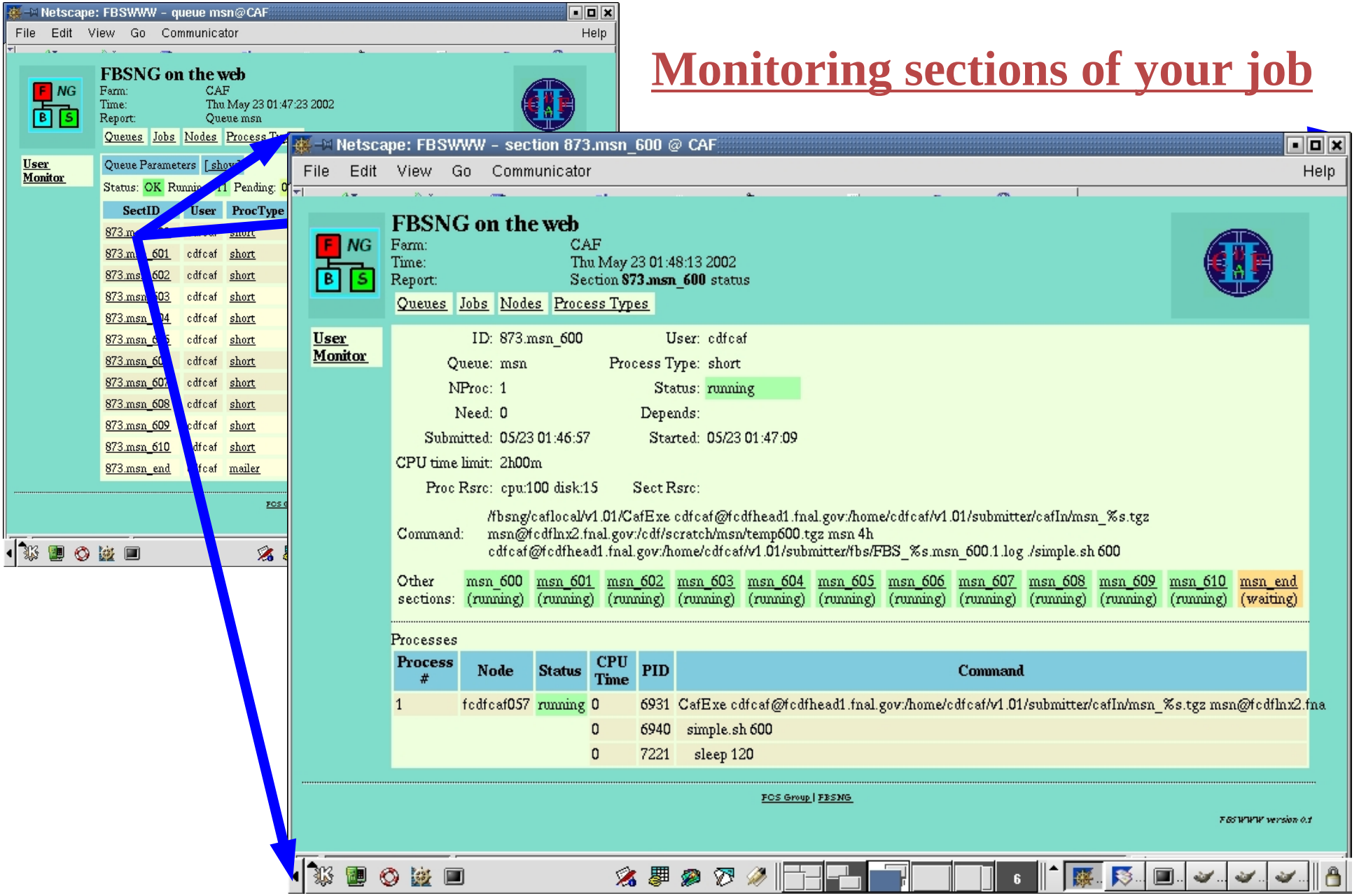
Queue Parameters [show]

Status: OK Running: 11 Pending: 0

SectID	User	ProcType	Status	Prio	NProc	Date/Time
873.msn_600	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:09
873.msn_601	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:09
873.msn_602	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:10
873.msn_603	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:10
873.msn_604	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:11
873.msn_605	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:11
873.msn_606	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:12
873.msn_607	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:12
873.msn_608	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:12
873.msn_609	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:13
873.msn_610	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:13
873.msn_end	cdfcaf	mailer	waiting	0	0/1	Submitted at 05/23 01:46:57

FCS Group | FBSNG

Monitoring sections of your job



CAF User Tools

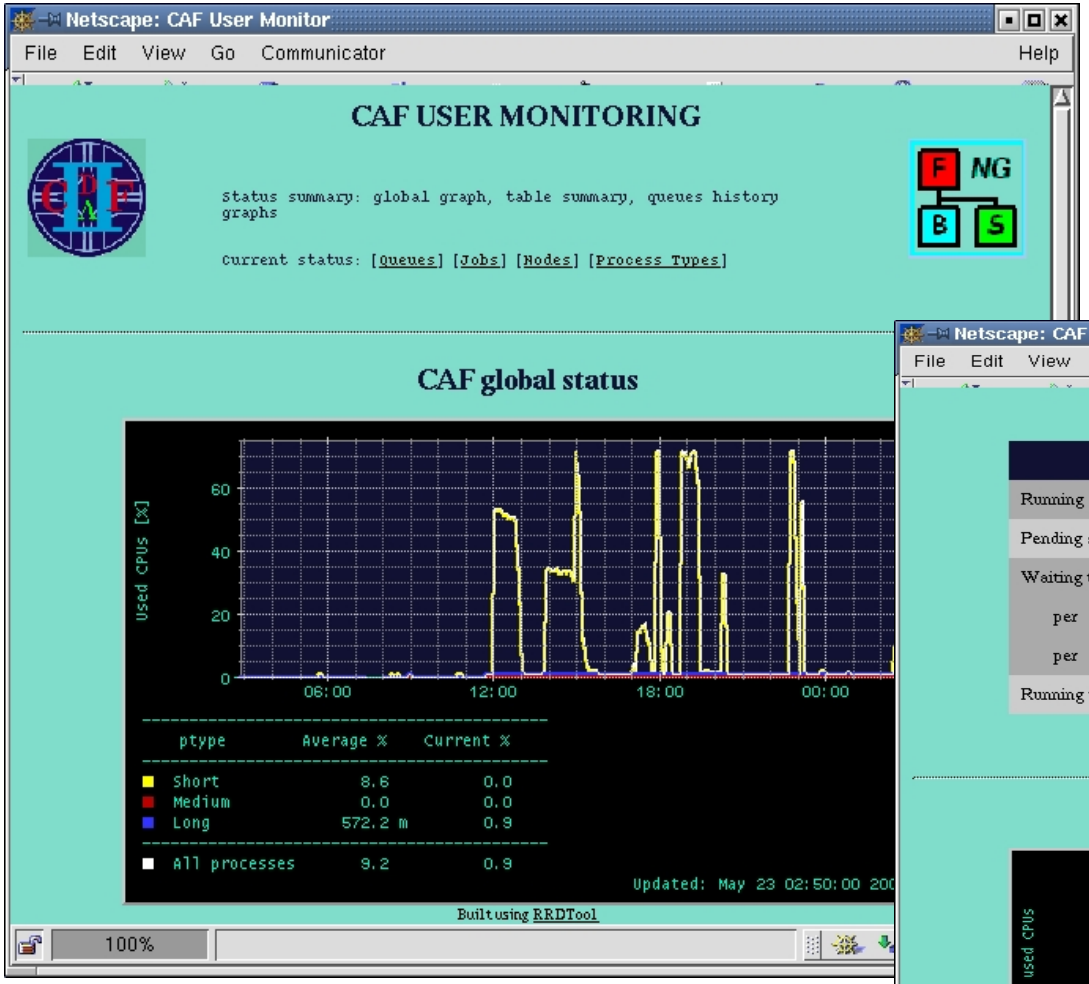
Job Control/Monitoring Utilities:

- **Get list/status of jobs in your queue**
  > cafjobs
- **Check progress of a section**
  > caflog JID sectionNumber
- **Kill a job or sections within a job**
  > cafkill JID [SectionRangeList]

Remote file listings:

- **Get listing in a section's 'relative path'**
  > cafdir JID sectionNumber [directory]
- **List files in a CAF node's absolute path**
  > cafhostdir hostname directory
- **Get full file listing for a dataset**
  > cafdataset datasetid

Monitoring CAF utilization



...by process type and queue

Netscape: CDF CAF Admin Monitoring page

File Edit View Go Communicator

CDF Central Analysis Facility

Admin Monitoring Page

Check the statistics to view

☐ ALL

☐ Load averages ☐ CPU usage ☐ Network throughput ☐ Memory usage

Check the nodes to view above statistics on:

Utility nodes

☐ ALL

☐ fcdfhead1 ☐ fcdfmon1

Servers

☐ ALL

☐ fcdfdata001 ☐ fcdfdata005

☐ fcdfdata004 ☐ fcdfdata006

☐ fcdfdata007

☐ fcdfdata008

Worker nodes

☐ ALL

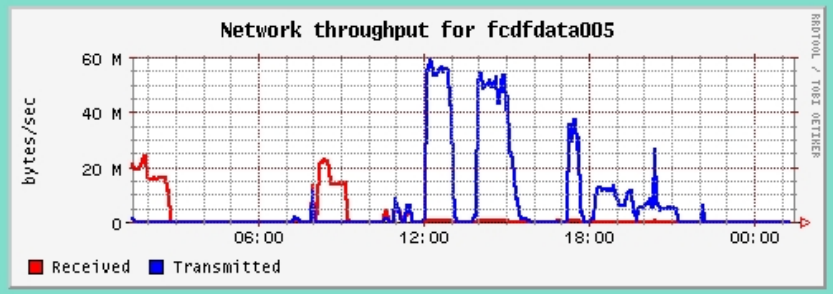
☐ fcdfcaf001 ☐ fcdfcaf020 ☐ fcdfcaf039

☐ fcdfcaf002 ☐ fcdfcaf021 ☐ fcdfcaf040

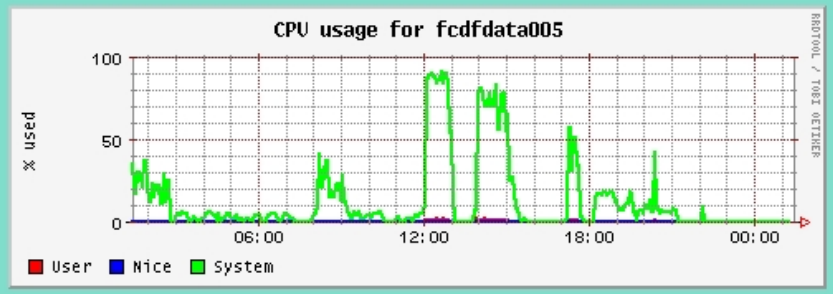
Netscape: CDF CAF Admin Monitoring Graphs

File Edit View Go Communicator Help

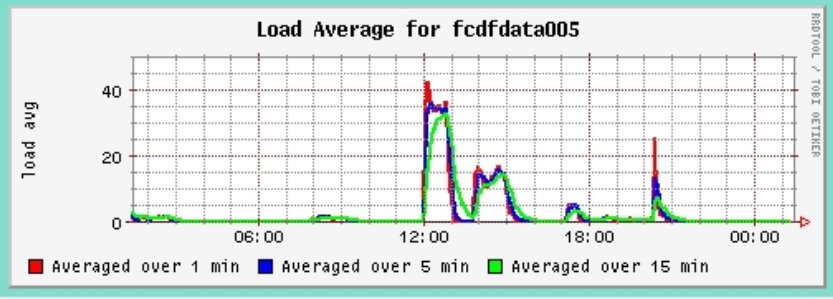
Network throughput



CPU usage



Load averages



Job Output Destinations

CAF ftp servers

Job and scratch space for each user:

- **Job space** (icaf/) - job output tarballs
  Specify *icaf:temp.tgz* for 'Output File Location' in CAF GUI
- **Scratch space** (scratch/) - data for subsequent job(s)
  (currently rcp, later kerberized rootd)

Both spaces accessible via kerberized ftp and ICAF tools:

> icaf_gftp (ICAF GUI)
> icaf_info (ICAF user-related info)
> icaf_ls/rm/get

Remote 'desktop'

Output can be sent:

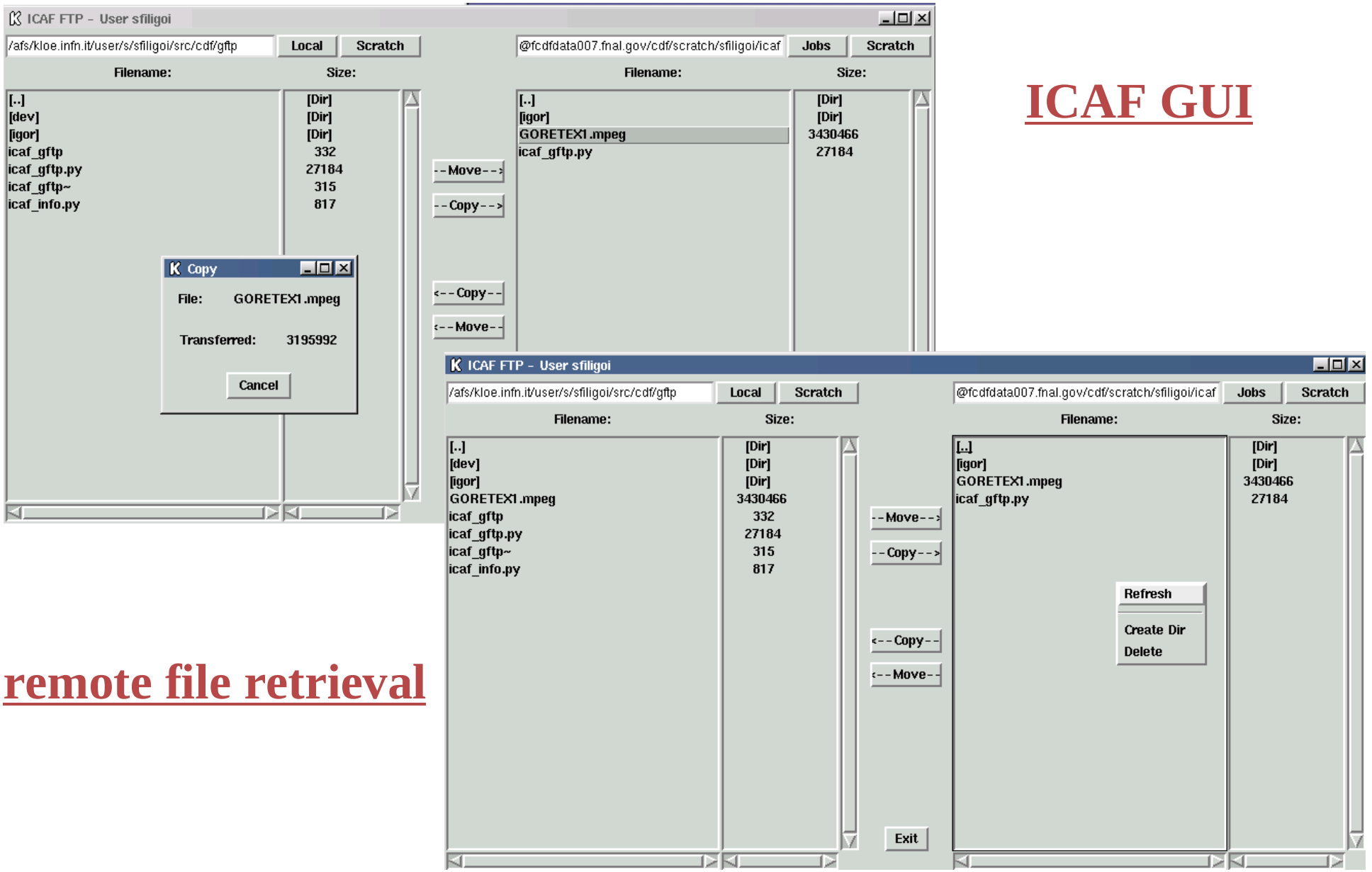
- directly back to desktop
- other remote machine

Any machine allowing incoming kerberized rcp can get CAF output

```
> cat ~johndoe/.k5login
johndoe@FNAL.GOV
caf/cdf/johndoe@FNAL.GOV
```

special CAF principal for each user

ICAF GUI



remote file retrieval

Data Access on the CAF

Data access methods - NFS & rootd

- **NFS - server data exported to each worker**

```
talk DHInput
    input file /cdf/data/fcdfdata005/jbot0c/*
exit
```

- **rootd - remote access to root files**

```
talk DHInput
    input file root://fcdfdata005.fnal.gov//export/data/jbot0c/*
exit
```

Available in DHInput after 4.4.0int6 (no extra work for 4.5.0)

CAF data servers and ftp servers run rootd

→ can run CAF job(s) on output of previous CAF job(s)

Data Management

- **No explicit Data Handling functionality for initial Stage1**
  → Secondary datasets disk-resident on CAF servers (fcdfdata001, fcdfdata004-006)
- **Datasets on fcdfsgi2 mirrored onto CAF servers**
  → Space managed by Strippers Club (C. Paus, et al)
- **Servers currently X% populated with datasets X,Y,Z**
- **Get server data listing using cafdataset/cafhostdir utilities:**

```
> cafdataset jbot0c
> cafhostdir fcdfdata005 /export/data/jbot0c
```

Interfacing to DH big part of evolving Stage 1 CAF!

- Initial Stage1: ~20 TB for static data, 9 TB for dCache development
- Plan: all disk eventually used for dCache/Enstore

Future Plans

Near term:

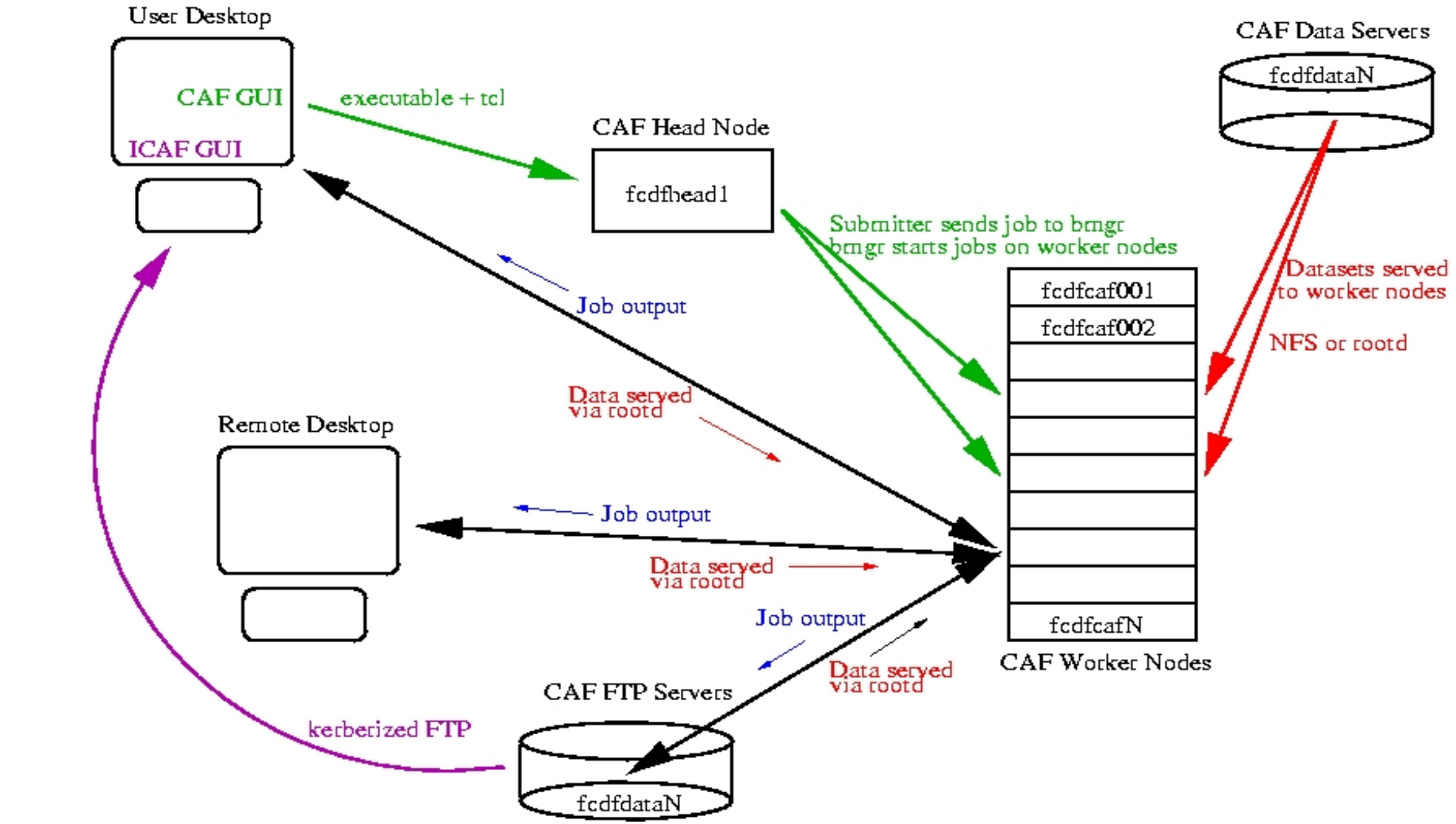
- **Gain operational experience!!**
  - → fix bugs, learn usage patterns & bottlenecks
- **Kerberized rootd**
- **Integrate Data Handling system**
- **Eliminate single points of failure**
- **Develop monitoring/alarms capability** (NGOP, RRD)
- **University ownership issues**
- **Database replicas**

FY03 and beyond:

- **Scalibility issues**
- **DCAF + GRID**

Summary/Conclusions

- **Paradigm shift in CDF computing**
- **CAF Design→Prototype→Stage1 ≤ 6 months**
- **CAF open business!**
  - → **Send account request to:**
    fkw@fnal.gov + msn@fnal.gov + thkim@fnal.gov
  - → **Join CDF_CAF and CDF_CAF_USER mailing lists**
  - → **Read the User's Guide**
    available at cdfcaf.fnal.gov (CAF home page)
  - → **Try it out & send CAF-related questions/problems to:**
    cdf_caf@fnal.gov


User Desktop
CAF GUI
ICAF GUI
executable + tcl
CAF Head Node
fcdfhead1
CAF Data Servers
fcdfdataN
Submitter sends job to bmgr
bmgr starts jobs on worker nodes
Datasets served
to worker nodes
NFS or rootd
Job output
Data served
via rootd
fcdfcaf001
fcdfcaf002
fcdfcafN
CAF Worker Nodes
Remote Desktop
Job output
Data served
via rootd
Job output
Data served
via rootd
CAF FTP Servers
kerberized FTP
fcdfdataN

CDF CAF File Server Benchmarking